PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2000-148783**
(43)Date of publication of application : **30.05.2000**

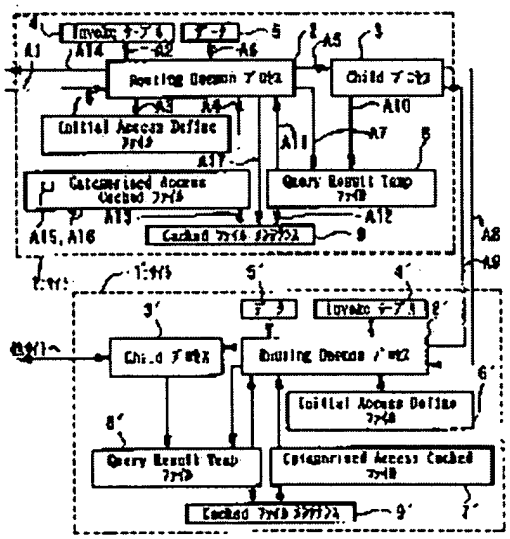
(51)Int.Cl. **G06F 17/30**
G06F 13/00

(21)Application number : **10-324249**
(22)Date of filing : **13.11.1998**

(71)Applicant : **NEC CORP**
(72)Inventor : **KIKUCHI SHINJI**

(54) DATA RETRIEVAL DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To widen the area of data retrieval using the internet and to make the retrieval efficient.
SOLUTION: When a site 1 receives a data retrieval request A1 from a user, a Routing Daemon process 2 actuates a Child process 3 by referring to data 5 in its device and makes another side 1' outputs a data retrieval request A8. This data retrieval request A8 includes address information A3 on a site to which the data retrieval request is always transferred and address information A4 on a site to which the most likelihood data retrieval request should be transferred from an Initial Access Define file 6. Transfer data A9 retrieved at the other side 1' are held in a Query Result Temp file 8 together with corresponding data A7 by the data 5, merged, and sent as transfer data A14 back to the user. Cached file maintenance 9 extracts the address information on the most likelihood site according to data contents transferred from the Query Result Temp file 8 and updates the Query Result Temp file 8.



(2)

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－148783
（P2000－148783A）
(43)公開日　平成12年５月30日(2000.5.30)

(51)Int.Cl.⁷		識別記号	FI			テーマコード（参考）
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/40	３１０Ｆ	５Ｂ０７５
	13/00	３５４		13/00	３５４Ｄ	５Ｂ０８９
				15/40	３１０Ｃ	

審査請求　有　　請求項の数11　ＯＬ　（全 10 頁）

(21)出願番号　特願平10－324249

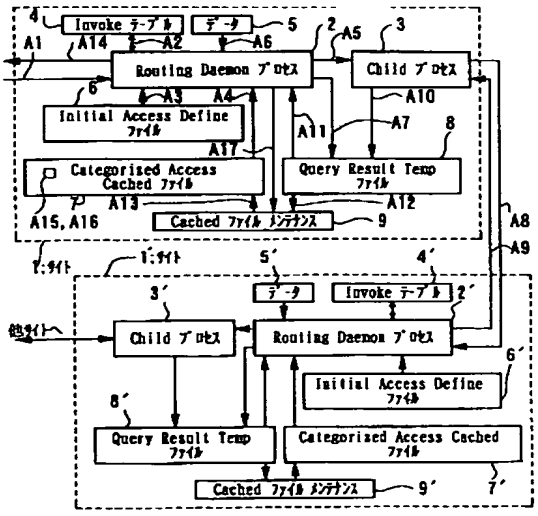
(22)出願日　平成10年11月13日(1998.11.13)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　菊地　伸治
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100108578
弁理士　高橋　詔男　（外３名）

Ｆターム(参考)　5B075 KK02 KK13 KK33 PQ05 PQ20
5B089 GA11 GB09 HA10 JA12 KC15
KC23 KC39 KC44

(54)【発明の名称】　データ検索装置

(57)【要約】
【課題】　インターネットを利用したデータ検索の広域化、高効率化を図る。
【解決手段】　サイト1が利用者からデータ検索要求A1を受信すると、RoutingDaemonプロセス2は、自装置内のデータ5を参照し、Childプロセス3を起動し、他のサイト1′にもデータ検索要求A8を出力させる。このデータ検索要求A8には、InitialAccessDefineファイル6からの、常にデータ検索要求を転送する サイトのアドレス情報A3と、最も確からしいデータ検索要求を転送すべきサイトのアドレス情報A4を含む。他のサイト1′　で検索された転送データA9は、データ5による該当データA7と共にQueryResultTempファイル8に保持され、マージされ、転送データA14として利用者に返される。Cachedファイルメンテナンス9は、QueryResultTemp ファイル8から転送されて きたデータ内容に応じて最も確からしいサイトのア ドレス情報を引き出し、QueryResultTempファイル8を更新する。



【特許請求の範囲】

【請求項１】　インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、

データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置からデータを受理し、前記自装置内に保持しているデータとマージしてデータ検索要求元に転送することを特徴とするデータ検索装置。

【請求項２】　インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、

データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置から検索データを受理する常駐のRouting Daemon プロセスと、

該Routing Daemon プロセスから起動され、複数の他のデータ転送装置とデータ転送に関する通信を行なうChild プロセスとを有することを特徴とするデータ検索装置。

【請求項３】　前記各データ転送装置は、新たなデータ検索要求を受けると、前記Routing Daemon プロセスによって起動識別子が登録されるInvokeテーブルを備え、Routing Daemon プロセスは、データ検索要求を受けた場合に、該Invokeテーブルを参照することにより、重複した処理を回避することを特徴とする請求項２記載のデータ検索装置。

【請求項４】　前記各データ転送装置は、前記Routing Daemon プロセスが、常に前記データ検索要求を転送する他のデータ転送装置のアドレスを格納したInitial Access Define ファイルを備えたことを特徴とする請求項２または請求項３記載のデータ検索装置。

【請求項５】　前記各データ転送装置は、前記Routing Daemon プロセスが、前記データ検索要求のデータ内容に応じて最も確からしいデータ検索要求を転送すべきデータ転送装置のアドレスを格納したCategorized Access Cached ファイルを備えたことを特徴とする請求項２ないし請求項４のいずれかに記載のデータ検索装置。

【請求項６】　前記各データ転送装置は、前記データ検索の結果、前記自装置内で参照されたデータおよび他のデータ転送装置から転送された検索データを一時的に保持するQuery Result Temp ファイルと、該Query Result Temp ファイルから転送されてきたデータ内容に応じて最も確からしいデータ転送先のデータ転送装置のアドレスを引き出し、この結果によって前記Categorized Access Cached ファイルを更新するCached ファイル　メンテナンスとを備えたことを特徴とする請求項５記載のデータ検索装置。

【請求項７】　前記最も確からしいデータ転送装置は、まず、検索条件の指定カテゴリに対し相対網羅度が最小のものを選択し、次に、相対新規度が最小のものを第１優先条件、前記相対網羅度が最小のものを第２優先条件としてソートして決定することを特徴とする請求項５または請求項６記載のデータ検索装置。

【請求項８】　前記最も確からしいデータ転送装置の決定方法は、計算上、扱うべきデータ転送装置の数を制限するサイト上限値設定手順と、保持している検索データ件数が多い順にソートして前記制限されたデータ転送装置の数に相当する順位迄のデータ転送装置を選択する該当サイト選択手順と、検索結果の対象をソートし、同じ対象に関する記述を持つ異なるデータ転送装置の数を前記対象の各々に対して数える該当オブジェクト選択手順と、前記選択された全データ転送装置に対して、各々、データ転送装置の数を変数Ｘとし、自身を含めたＸ台のデータ転送装置に含まれる前記対象の度数に関するヒストグラムを作成するヒストグラム作成手順と、該作成されたヒストグラムを基に網羅度および新規度を計算する網羅度・新規度算出手順と、前記網羅度の小さい順にソートしてデータ転送装置を並び替え、小さい順に前記相対網羅度を付与する相対網羅度付与手順と、前記新規度の小さい順にソートしてデータ転送装置を並べ替え、小さい順に前記相対度を付与する相対新規度付与手順と、該付与された相対網羅度，相対新規度と、前記Categorized Access Cached ファイルで管理されている現在の相対網羅度,相対新規度との相対平均をとり、前記Categorized Access Cached ファイルに書き込む更新手順とを含むことを特徴とする請求項７記載のデータ検索装置。

【請求項９】　前記Child プロセスは、起動の際、データ検索要求を転送すべき他のデータ転送装置のアドレス情報および他のデータ転送装置から転送されてきたデータを一時的に保持すべき前記Query Result Temp ファイルの名称情報を前記Routing Daemon プロセスから引数として受け取ることを特徴とする請求項２ないし請求項８のいずれかに記載のデータ検索装置。

【請求項１０】　前記データ検索要求は、転送される際に、通過するデータ転送装置のName特定情報および累積Hop 数が内部にシーケンス状に記録されることを特徴とする請求項１ないし請求項９のいずれかに記載のデータ検索装置。

【請求項１１】　前記データ検索要求，Invoke テーブル，Initial AccessDefine ファイル，Categorized Access Cached ファイルおよびQuery Result Temp ファイルBNF (Backus Normal Form) で定義される記述形式で表されることを特徴とする請求項１ないし請求項１０のいずれかに記載のデータ検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、インターネット上

のデータ検索要求に対するデータ転送装置に関する。

【０００２】

【従来の技術】従来のこの種の転送方式を図４に示す。この方式は、米国論文（M. A. sheldon, A. Duda, R. Weiss, D. K. Gifford:Discover:A Resource Discovery System based onContent Routing, The 3rd International world-wide web Conference April10-14. 1995 (http//www.igd.de/www/www95/papers/82/html-files/discover.html) (1995).）に記されたものである。

【０００３】この従来方式は、利用者に対してサーバの役割を果たすHttpd プロセス１１，ルーティング機能を提供する複数のコンテントルータ１２等，各コンテントルータが参照し、ルーティングに関する情報を管理する複数のルーティングデータベース１３等，各コンテントルータが参照し、利用者に正しい質問入力を促進するための複数のRefinement データベース１４等，検索機能を提供する複数のサーチモジュール１６，ドキュメントデータを管理するためのシステムである複数のWAIS プロセス１７および各WAIS プロセス１７が参照する複数のWAIS データベース１８を含んでいる。

【０００４】利用者がWeb プラウザ等を介して、Httpd プロセス１１にドキュメントの検索要求Ｂ１を発行すると、Httpd プロセス１１は、ルーティング機能を提供し、直接関連しているコンテントルータ１２に検索要求Ｂ１２を発行する。このコンテントルータ１２は、通常、Httpd プロセス１１のバックプロセスとして機能する。

【０００５】その後、コンテントルータ１２は、検索要求Ｂ１で指定され、検索要求Ｂ２に含まれる質問語彙妥当か否かを判定するため、Refinement データベース１４上の正しい質問データＢ３を参照し、妥当な表現に修正する。

【０００６】次に、コンテントルータ１２は、ルーティングデータベース１３に問い合わせ、質問語彙に対応して、該当する検索結果を応答できる、もしくはその仲立ちができる他のコンテントルータ１２′，１２″の位置情報Ｂ４を得る。この情報はルーティングデータベース１３の内部ではコンテンツラベル１５として管理されている。また他のルーティングデータベース１３′，１３″でも同様にコンテンツラベル１５′，１５″が存在している。

【０００７】次に、コンテントルータ１２は、コンテンツラベル１５から得た位置情報Ｂ４が指し示す他のコンテントルータ１２′，１２″に、検索要求Ｂ２と等価な検索要求Ｂ５，Ｂ５′を転送する。

【０００８】他のコンテントルータ１２′，１２″においても、コンテントルータ１２と全く同様の処理を行なう。その際、もしこのコンテントルータ１２′，１２″がルーティングの際に、最終位置にある場合、直接関連するサーチモジュール１６，１６′に検索要求Ｂ５，Ｂ５′と等価な検索要求Ｂ６，Ｂ６′を転送する。　サーチモジュール１６，１６′が検索要求Ｂ６，Ｂ６′を受けると、この検索要求Ｂ６，Ｂ６′と等価な検索要求Ｂ７，Ｂ７′を配下のドキュメントデータを管理するWAIS プロセス１７，１７′に発行する。それぞれのWAIS プロセス１７，１７′は検索要求Ｂ１で指定された質問語彙を含むＢ９，Ｂ９′を各WAIS データベース１８，１８′に発行し、関連するドキュメントの位置情報Ｂ１０，Ｂ１０′をWAIS プロセス１７，１７′に返す。WAIS プロセス１７，１７′は、対応するサーチモジュール１６，１６′に検索要求Ｂ７，Ｂ７′の応答として、ドキュメントの位置情報Ｂ１０，Ｂ１０′と等価な位置情報Ｂ１１，Ｂ１１′を返す。

【０００９】同様にサーチモジュール１６，１６′は対応するコンテントルータ１２′，１２″に検索要求Ｂ６，Ｂ６′の応答として、ドキュメントの位置情報Ｂ１１，Ｂ１１′と等価な位置情報Ｂ１２，Ｂ１２′を返す。その後コンテントルータ１２′，１２″は、起点となるコンテントルータ１２に検索要求Ｂ５，Ｂ５″の応答として、ドキュメントの位置情報１２，Ｂ１２′と等価な位置情報Ｂ１３，Ｂ１３′を返す。

【００１０】起点となるコンテントルータ１２は、配下のコンテントルータ１２′，１２″から検索結果であるドキュメントの位置情報Ｂ１３，Ｂ１３′を全て受けとると、それらを合成し最終的な位置情報Ｂ１４としてHttpd プロセス１１に戻す。その後、利用者が発行したドキュメントの検索要求Ｂ１に対応したドキュメントの位置情報Ｂ１５が利用者のWeb プラウザ等上に表示されることになる。

【００１１】コンテンツラベル１５，１５′，１５″を更新する場合は、コンテントルータ１２，１２′，１２″は検索条件を限定しない検索要求Ｂ１６，Ｂ１６′を発行する。その後、前述の手順に応じて、WAIS データベース １８，１８′上に管理された全ドキュメントの位置情報Ｂ１７，Ｂ１７′が返されることになる。全ドキュメントの位置情報Ｂ１７，Ｂ１７′の内容はコンテンツラベル１５，１５′，１５″を更新し得るだけの内容を持ち、情報内容を持って、コンテントルータ１２，１２′，１２″はコンテンツラベル１５，１５′，１５″を更新する。

【００１２】

【発明が解決しようとする課題】上述した従来の方式２″では、図４に示した様にWAIS プロセス１７，１７′を使っているため、ドキュメント検索の際、語彙検索が中心となる。また、基本的にWAIS データベース１８，１８′で管理されるコンテンツデータを対象としており、一般的なレガシーデータを対象とはしていないという第１の問題点がある。

【００１３】また、コンテンツラベル１５，１５′，１５″を更新する手段を用意してはいるが、検索の際には

最適化した検索が行われる機構が特になく、決められたコンテントルータ１２′，１２″にのみ、常に検索要求をルーティングすることしか出来ないという第２の問題点がある。

【００１４】また、仮に、２つのルーティングデータベース１３′，１３″内の各々のコンテンツラベル１５′，１５″上に、互いに相手の位置情報が登録されている場合には、検索が２重に実施されることになり検索上のロスが発生するという第３の問題点がある。

【００１５】従って、本発明の目的は、近年、益々、重要になって来ているインターネット上のドキュメントを含むデータベースの検索を出来るだけ広域にかつ無駄なく、実施出来るデータ転送装置を提供することにある。

【００１６】また、本発明の他の目的は、利用者の欲している情報をできるだけ定量的に評価して効率を高めることができるデータ検索装置を提供することにある。

【００１７】

【課題を解決するための手段】そこで、本発明の第１のデータ検索装置は、インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置からデータを受理し、前記自装置内に保持しているデータとマージしてデータ検索要求元に転送することを特徴とする。また、本発明の第２のデータ検索装置は、インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置から検索データを受理する常駐のRouting Daemon プロセスと、該Routing Daemon プロセスから起動され、複数の他のデータ転送装置とデータ転送に関する通信を行なうChildプロセスとを有することを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、新たなデータ検索要求を受けると、前記Routing Daemon プロセスによって起動識別子が登録されるInvoke テーブルを備え、Routing Daemon プロセスは、データ検索要求を受けた場合に、該Invoke テーブルを参照することにより、重複した処理を回避することを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、前記Routing Daemon プロセスが、常に前記データ検索要求を転送する他のデータ転送装置のアドレスを格納したInitial Access Define ファイルを備えたことを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、前記Routing Daemon プロセスが、前記データ検索要求のデータ内容に応じて最も確からしいデータ検索要求を転送すべきデータ転送装置のアドレスを格納したCategorized Access Cached ファイルを備えたことを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、前記データ検索の結果、前記自装置内で参照されたデータおよび他のデータ転送装置から転送された検索データを一時的に保持するQuery Result Tempファイルと、該Query Result Temp ファイルから転送されてきたデータ内容に応じて最も確からしいデータ転送先のデータ転送装置のアドレスを引き出し、この結果によって前記Categorized Access Cached ファイルを更新するCached ファイル　メンテナンスとを備えたことを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記最も確からしいデータ転送装置は、まず、検索条件の指定カテゴリに対し相対網羅度が最小のものを選択し、次に、相対新規度が最小のものを第１優先条件、前記相対網羅度が最小のものを第２優先条件としてソートして決定することを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記最も確からしいデータ転送装置の決定方法は、計算上、扱うべきデータ転送装置の数を制限するサイト上限値設定手順と、保持している検索データ件数が多い順にソートして前記制限されたデータ転送装置の数に相当する順位迄のデータ転送装置を選択する該当サイト選択手順と、検索結果の対象をソートし、同じ対象に関する記述を持つ異なるデータ転送装置の数を前記対象の各々に対して数える該当オブジェクト選択手順と、前記選択された全データ転送装置に対して、各々、データ転送装置の数を変数Ｘとし、自身を含めたＸ台のデータ転送装置に含まれる前記対象の度数に関するヒストグラムを作成するヒストグラム作成手順と、該作成されたヒストグラムを基に網羅度および新規度を計算する網羅度・新規度算出手順と、前記網羅度の小さい順にソートしてデータ転送装置を並べ替え、小さい順に前記相対網羅度を付与する相対網羅度付与手順と、前記新規度の小さい順にソートしてデータ転送装置を並べ替え、小さい順に前記相対新規度を付与する相対新規度付与手順と、該付与された相対網羅度，相対新規度と、前記Categorized Access Cached ファイルで管理されている現在の相対網羅度,相対新規度との相対平均をとり、前記Categorized Access Cached ファイルに書き込む更新手順とを含むことを特徴とする。

【００１８】

【発明の実施の形態】次に、本発明の実施の形態について説明する。図１は、本発明のデータ内容に応じたルーティング方式を採用したデータ転送装置の実施例を示す。本実施例では、データ転送装置をサイト（Site）と称し、図１には２つのサイト１と１′が示されている。

【００１９】サイト１にはデータ検索要求を受けると、自装置内に保持しているデータを参照すると共に他サイ

ト１′へ同一のデータ検索要求を転送・発行し，その結果、他サイト１′内に保持しているデータを受理する常駐プログラムのRouting Daemon プロセス２と、Routing Daemon プロセス２から起動され、他の、サイト１′等の複数のデータ転送装置と実際のデータ転送に関する通信を行なうChild プロセス３と、データ検索要求の処理状況を管理・記録するInvoke テーブル４と、サイト内で固有に保持されているデータ５と、Routing Daemon プロセス２が常にデータ検索要求を転送・発行するサイトのアドレスを記したInitial Access Define ファイル６と、後述の該当データマイニング方式により更新され、データ検索要求を転送・発行する際に最も確からしいサイトのアドレスを得るのに使用するCategorizedAcess Cached ファイル７と、データ検索要求によるデータ５への参照結果であるデータ、並びに上述のデータ検索要求の結果、転送されて来たサイト１′内に保持されているデータ５′を一時的に保持・記録するQuery Result Temp ファイル８と、データマイニング方式により、Query Result Temp ファイル８からデータ検索要求の条件に応じて最も確からしいデータ転送先のサイトのアドレスを引き出し、その結果をもってCategorized Access Cached ファイル７を更新・維持管理するCached ファイル　メンテナンス９を含んでいる。

【００２０】Invoke テーブル４は、各サイトに１つ設定されるものであり、サイト間でやりとりされるデータ検索要求の処理状況を管理・記録するものである。RoutingDaemon プロセス２がデータ検索要求A_1を受理すると、Invoke テーブル４をアクセスし、該当する起動識別子（以下InvokeID ）を含んだ情報A_2が存在するか否かを確認する。このInvokeID を含んだ情報A_2を初めて扱う場合は、InvokeIDを含んだ情報A_2をテーブル４に書き込むことで登録する。それに対して、サイト１がサイト１′内のChild ３′から、既に同一のInvokeID に相当するデータ検索要求A_1を受けている場合は、そのInvokeID を含んだ情報は、登録済みになっているので、新たに同じデータ検索要求を受けた場合、Routing Daemon プロセス２が処理を無視する、もしくは拒否する等を行ない、２重に同一の処理が実施されることを防止する。

【００２１】Invoke テーブル４は、以下のBNF(Backus Normal Form)で定義される記述形式を持ち、同一のデータ検索要求は容易に判定出来る。

【００２２】

【数１】

```
InvokeTable ::= ⟨originl_pass⟩ ⟨invokeID⟩;
⟨originl_pass⟩ ::= ⟨identifier⟩ ;
⟨identifier⟩ ::= ⟨ip_string⟩ | ⟨dns_string⟩ | ⟨other⟩ ;
⟨ip_string⟩ ::= {⟨number⟩ 3} {'.' ⟨number⟩ 3} 3;
⟨dns_string⟩ ::= {⟨word⟩'.'} {⟨word⟩'.'} *;
⟨other⟩ ::= ⟨word⟩ *;
⟨string⟩ ::= ⟨character⟩ | ⟨character⟩ ⟨character⟩ *;
⟨word⟩ ::= ⟨alphabetic⟩ {⟨alphabetic⟩ *} ;
⟨character⟩ ::= ⟨alphabetic⟩ | ⟨number⟩ | ⟨special⟩ ;
⟨number⟩ ::=1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0;
⟨special⟩ ::= ⟨ | ⟩ | = | + | - | * | / | & | ^ | ~ | _ | @ | $ | % | : | . | ! | ?;
⟨invokeID⟩ ::= ⟨number⟩ 15;
```

【００２３】また、データ検索要求A_1は、以下のBNFで定義される記述形式を持ち、転送の際に通過サイトのName 特定情報、並びに累積Hop 数が内部に記憶される。尚、通過サイトのName 特定情報は、順序が把握出来るようにシーケンス状に記録される。

【００２４】

【数２】

```
QueryText ::=⟨originl_pass⟩⟨invokeID⟩⟨query⟩⟨hop_number⟩⟨pass_list⟩⟨local_pass⟩;
⟨hop_number⟩ ::= ⟨number⟩ 3;
⟨pass_list⟩ ::=' | ' {⟨remote_pass⟩' | '} *;
⟨remote_pass⟩ ::= ⟨identifier⟩ ;
⟨local_pass⟩ ::= ⟨identifier⟩ ;
⟨originl_pass⟩ ::= ⟨identifier⟩ ;
⟨identifier⟩ ::= ⟨ip_string⟩ | ⟨dns_string⟩ | ⟨other⟩ ;
⟨ip_string⟩ ::= {⟨number⟩ 3} {' .' ⟨number⟩ 3} 3;
⟨dns_string⟩ ::= {⟨word⟩' .'} {⟨word⟩' .'} *;
⟨other⟩ ::=- ⟨word⟩ *;
⟨query⟩ ::= ⟨string⟩ ;
⟨string⟩ ::= ⟨character⟩ | ⟨character⟩ ⟨character⟩ *;
⟨word⟩ ::= ⟨alphabetic⟩ {⟨alphabetic⟩ *} ;
⟨character⟩ ::= ⟨alphabetic⟩ | ⟨number⟩ ⟨special⟩ ;
⟨number⟩ ::=1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0;
⟨special⟩ ::= ⟨ | ⟩ | = | + | - | * | / | & | ^ | ~ | _ | @ | $ | % | : | . | ! | ?;
⟨invokeID⟩ ::= ⟨number⟩ 15 ;
```

【００２５】Routing Daemon プロセス２は、その後、categorized Access Cached ファイル７をアクセスし、先のデータ検索要求A_1を転送・発行するべきサイトのアドレス情報A_4を得ると共に、Initial Access Define ファイル６もアクセスし、同様に先のデータ検索要求A_1を転送・発行するサイトのアドレス情報A_3を得る。Initial Access Define ファイル６には、Routing Daemon プロセス２が常にデータ検索要求A_1を転送・発行するサイトのアドレスが記されているのに対して、Categorized Access Cached ファイル７には、データ検索要求A_1で指定された検索条件としてのカテゴリに応じて、データ検索要求を転送・発行する際に最も確からしいサイトのアドレスが記されている。

【００２６】その際の最も確からしいサイトの決定手順は、図２に記されたようになる。ステップＳ１として、まず、後述の相対網羅度A_{15}が最小のものを、利用者が検索条件として指定したカテゴリに対し、最も有効と思われるサイトとして選択する。相対網羅度A_{15}が最小ということは、もっとも多くの該当するデータを保持しているサイトを意味する。

【００２７】次に、ステップＳ２では、利用者が検索条件として、指定したカテゴリに対し、以下の優先条件でソートし、複数のサイトを規定数分、選択する。

第一優先条件）相対新規度Ａ１６が最小のもの

第二優先条件）相対網羅度Ａ１５が最小のもの

この条件設定の理由は、可能な限り、どのサイトも扱っていないものを検索対象にするためである。

【００２８】前述の様にInitial Access Define ファイル６には、サイト１内のChild プロセス３が、常にデータ検索要求A_1を転送・発行するべきアドレス情報Ａ３が指定されているが、このInitial Access Define ファイル６は、以下の様なBNF で表される記述形式を持つ。

【００２９】

【数３】

```
InitialAccessDefinefile::= {⟨indentifier⟩' | '} *;
⟨identifier⟩ ::= ⟨ip_string⟩ | ⟨dns_string⟩ | ⟨other⟩ ;
⟨ip_string⟩ ::= {⟨number⟩ 3} {' .' ⟨number⟩ 3} 3;
⟨dns_string⟩ ::= {⟨word⟩' .'} {⟨word⟩' .'} *;
⟨other⟩ ::=- ⟨word⟩ *;
```

【００３０】次に、Routing Daemon プロセス２は、サイトデータ検索要求Ａ１を転送・発行するべきサイトのアドレス情報Ａ３，Ａ４にChild プロセス３を該当数分だけ立ち上げる。起動の際、各々のChild サイトプロセス３はデータ検索要求A_1を転送・発行するべきアドレス情報Ａ３，Ａ４，並びにデータ検索要求Ａ１の結果、転送されて来た内に保持されているデータ５′を、一時的に保持・記録するQuery Result Temp ファイル８の名称情報を引数Ａ５として受け取る。任意のChild プロセス３がサイト１′に該データ検索要求Ａ１と等価なデータ検索要求Ａ８を転送・発行する際は、サイト１′内のRouting Daemon プロセス２′が受理する。その際、データ検索要求Ａ８にはサイト１のアドレス情報が付加され、データ検索要求Ａ８内部のHop数が１つ追加される。

【００３１】次に、Routing Daemon プロセス２は、自サイト内で管理しているデータ５からデータ検索要求Ａ１に応じたデータのサブセットＡ６を受けとり、一時的に保持・記録する為、のQuery Result Temp ファイル８に該当データＡ７として書き込む。Query Result Temp ファイル８は、以下のBNFで表される記述形式を持ち、プロセス３ごとに１つ作成される。

【００３２】

【数４】

```
QueryResultTempFile::= ⟨header⟩ {⟨URLdescription⟩ *} ;
⟨header⟩ ::= ⟨querytext⟩ ⟨amount⟩ ;
⟨amount⟩ ::= ⟨number⟩ 6;
⟨URLdescription⟩ ::= ⟨URL⟩ ⟨object_identifier⟩ ⟨description⟩ ;
⟨URL⟩ ::=' http://' {⟨ip_string⟩ | ⟨dns_string⟩}' /' ⟨string⟩ ;
⟨object_identifier⟩ ::= ⟨string⟩ ;
⟨description⟩ ::= ⟨string⟩ ;
```

【００３３】サイト１′内でもサイト１と同様に、Routin Daemon プロセス２′が同じ処理を行ない、サイト１′内で管理しているデータ５′からデータ検索要求Ａ８に応じたデータのサブセットＡ６を引き出し、転送データＡ９としてサイト内のChild プロセス３に戻す。Child プロセス３は転送データＡ９を受理すると、引数Ａ５で指定されたQuery Result Temp ファイル８に該当データＡ１０として書き込む。この処理は、起動されたChild プロセス３数分だけ実施される。

【００３４】次に、Routing Daemon プロセス２は全てのChild プロセス３が該当データＡ１０を書き込んだことを確認すると、全てのQuery Result Temp ファイル８をマージし、そこから該当データＡ１１を読み出す。その後、Routing Daemon プロセス２は当該データＡ１１をデータ検索要求Ａ１に応じた転送データＡ１４として、データ検索要求元に転送する。以上で、一連のデータ検索要求Ａ１は完了する。Routing Daemon プロセス２は、不要となったQuery Result Temp ファイル８を消去する際に、Cached ファイル　メンテナンス９へ起動要求１７を送付する。Cached ファイル　メンテナンス９は起動すると、データ検索結果を格納したQuery Result Temp ファイル８から、Categorized Access Cached ファイル７上の相対網羅度Ａ１５，相対新規度Ａ１６を更新するためのパラメータデータＡ１２を引き出し、これを基に更新元データＡ１３を作成し、Categorized Access Cached ファイル７上の現相対網羅度Ａ１５、現相対新規度Ａ１６を更新する。前述の様にCategorized Access Cached ファイル７とは、データ検索要求Ａ１，Ａ８を転送した結果、得られる一時的なデータ検索格納先であるQuery Result Temp ファイル８から、事前に定義した当該データのマイニング方式に応じて、各サイトの評価を行ない、そのアドレス情報を管理するものである。Categorized Access Cached ファイル７により、Routing Daemon プロセス２は、データ検索要求Ａ１，Ａ８を転送・発行する際に、検索条件として指定されるカテゴリに対し最も有効と思われるサイトのみに転送・発行先を絞り込むことが出来るので、全サイトへ単純にブロードキャスト転送を行なうよりは、遥かに効果的なルーティング転送・発行処理を実現することが出来る。Categorized Access Cached ファイル７は以下のBNF で表される記述形式を持つ。

【００３５】

【数５】

```
CategorizedAssessCachedfile::= {⟨registry description⟩' |'} *;
⟨registry description⟩ ::= ⟨category identifier⟩ ⟨identifier⟩ ⟨metrix description⟩ ;
⟨category identifier⟩ ::= ⟨string⟩ ;
⟨identifier⟩ ::= ⟨ip_string⟩ | ⟨dns_string⟩ | ⟨other⟩ ;
⟨ip_string⟩ ::= {⟨number⟩ 3} {' .' ⟨number⟩ 3} 3;
⟨dns_string⟩ ::= {⟨word⟩' .'} {⟨word⟩' .' } *;
⟨other⟩ ::= ⟨word⟩ *;
⟨metrix description⟩ ::= ⟨relative covering degree⟩ ⟨relative strange degree⟩ ⟨counter⟩ ;
⟨relative covering degree⟩ ::= ⟨number⟩ 6;
⟨relative strange degree⟩ ::= ⟨number⟩ 6;
⟨counter⟩ ::= ⟨number⟩ 6;
```

【００３６】Categorized Access Cached ファイル７上で管理される相対網羅度Ａ１５，相対新規度Ａ１６は、以下に記されるものである。

【００３７】１）相対網羅度Ａ１５

Cached ファイル　メンテナンス９は当該データのマイニング方式としてサイト１′を始めとする関連する複数のサイトがデータ検索要求Ａ１，Ａ８に相当する任意データ検索要求を受理した結果として戻す転送データＡ９，Ａ１４に対し、各サイトがどの程度、検索条件である指定カテゴリに該当するものを含んでいるかを計算し、更新元データＡ１３を求める。その後、その計算値である更新元データＡ１３で、Categorized Access Cached ファイル７上で管理されるサイト，カテゴリ毎に記録している相対網羅度Ａ１５の値を更新する。相対網羅度Ａ１５の評価に関しては後述する。

【００３８】２）相対新規度Ａ１６

Cached ファイル　メンテナンス９は、前述の当該データのマイニング方式として、サイト１′を始めとする関連する複数のサイトがデータ検索要求Ａ１，Ａ８に相当する任意データ検索要求を受理した結果として戻す転送

データＡ９．Ａ１４に対し、各サイトがどの程度、検索条件である指定カテゴリに該当するもので独自かつ固有なものを含んでいるかを計算し、更新元データＡ１３を求める。その後その計算値である更新元データＡ１３で、Categorized Access Cached ファイル７上で管理される毎に記録している相対新規度Ａ１６の値を更新する。相対新規度Ａ１６の評価式に関しては後述する。

【００３９】次に、当該データのマイニング方式について説明する。図３は、該当データのマイニング方式のフローチャートである。ステップＳ１では、初期処理として、計算上、扱うべきサイト数を制限する。これは転送データＡ９を戻すサイトは基本的に不特定多数であることから配慮されている。これを「サイト上限値ステップ」と呼ぶ。

【００４０】ステップＳ２はChild プロセス３毎に作成されるQuery Result Temp ファイル８から、<header>記述を集める処理である。<header>記述内部には、データ件数を意味する<amount>が記されている。

【００４１】ステップＳ３では，<amount>記述中の値で大きい順にソートして<header>記述を並べ替え、ステップＳ１で制限されたサイト数に相当する順位のもの迄の複数サイトを処理の対象として選択する。これを「該当選択ステップ」と呼ぶ。

【００４２】ステップＳ４では、選択した複数サイトに対応するQuery Result Temp ファイル８中の<URLdescription>記述を全て取り出し、１つの<URLdescription>記述を１レコードと見なして、単純に連結する。なお、<URLdescription>記述の内部には、検索結果の対象を意味する<object identifier>記述が記されており、各々が１つの検索対象と見なされる。

【００４３】ステップＳ５では，<object identifier>記述でソートし、同じ<object identifier>記述を持つ異なるサイトの数を求める。以上を異なる<object identifier>記述毎全てにわたり実施する（ステップＳ６）。これを「該当object 選択ステップ」と呼ぶ。

【００４４】ステップＳ７では、ステップＳ３で選択した全サイトに対して（ステップＳ１６）、各々、サイト数を変数Ｘとし、自身を含めたＸ台のサイトに含まれる<object identifier>記述の度数ｍ（Ｘ）に関するヒストグラムを作成する。このステップＳ６では、ステップＳ３で選択した全サイトにわたり実施する。これを「ヒストグラム作成ステップ」と呼ぶ。

【００４５】ステップＳ８では、Ｓ７で求めたヒストグラムを基に網羅度及び新規度を以下の様に計算する。これを「網羅度・新規度算出ステップ」と呼ぶ。ステップＳ８も、ステップ７と同様に、ステップＳ３で選択した全サイトに対して行なう（ステップＳ１７）。

【００４６】

【数６】

$$網羅度＝\sum_{x}\{m(X)*X\}$$

$$新規度＝\sum_{x}\{m(X)/X\}$$

【００４７】ステップＳ９では、ステップＳ８で求めた網羅度の小さい順に選択したサイトをソートして、並べ替える。その後、小さい順に相対網羅度を１位，２位，３位と付与する。これを「相対網羅度付与ステップ」と呼ぶ。

【００４８】ステップＳ１０では、ステップＳ８で求めた新規度の小さい順に選択したサイトをソートして、並べ替える。その後、小さい順に相対新規度を１位，２位，３位と付与する。これを「相対新規度付与ステップ」と呼ぶ。

【００４９】ステップＳ１１では、ステップＳ９並びにステップＳ１０で求めた相対網羅度Ａ１５、相対新規度Ａ１６を更新値Ａ１３とする。その後、Categorized AccessCached ファイル７上でサイト，カテゴリ毎に管理されている現在の相対網羅度Ａ１５、現在の相対新規度Ａ１６を読み出す。これを「既存データ読みだしステップ」と呼ぶ。

【００５０】ステップＳ１３では、以下の処理を行なう。ステップＳ１１で現在の相対網羅度Ａ１５，現在の相対新規度Ａ１６が読み出せる場合、該当する各々と更新値Ａ１３との相加平均を取り、新たな相対網羅度Ａ１５、相対新規度Ａ１６を求める。その際、相加平均の対象数を表す<Counter>記述値に１を加える。その後、Categorized Access Cached ファイル７に新たに算出した相対網羅度Ａ１５、並びに新たに算出した相対新規度Ａ１６並びに<Counter>記述を書き込み、更新する。

【００５１】ステップＳ１１で現在の相対網羅度Ａ１５、現在の相対新規度Ａ１６が読み出せない場合は、ステップＳ９で求めた相対網羅度Ａ１５、ステップＳ９で求めた相対新規度Ａ１６、並びに該<Counter>記述を１として、Categorized Access Cached ファイル７に書き込み、更新する（ステップＳ１４）。これを「更新ステップ」と呼ぶ。

【００５２】ステップＳ１１並びにＳ１３，Ｓ１４は、ステップＳ３で選択した全サイト数回、処理を行う（ステップＳ１５）。その後、規定回数分の処理を終えたならば、該当データのマイニング方式全体の処理を終了する。

【００５３】

【発明の効果】本発明によれば、従来のように、語彙検索のみを対象とはしていないため、数値表現を含んだデータの検索が可能であり、一般的なレガシーデータをもその対象とすることが可能であるという第１の効果を有する。

【００５４】また、本発明では、データのマイニング方式を採用しているので、動的にルーティングを実施出来、その結果、新たなサイトが登録され、それが該当デ

ータのマイニング方式の評価に対して妥当なデータを戻す場合は、新たなルーティング対象として、これを取り込むことが可能であるという第２の効果を有する。

【００５５】さらに、本発明は、データ検索要求の起動状況を管理・記録するInvoke テーブルが実装されているので、検索が２重になることはないという第３の効果も有する。

【図面の簡単な説明】

【図１】　本発明のデータ検索装置の一実施例の構成図。

【図２】　本発明におけるルーティング先を決定する際の手順を示すフローチャート。

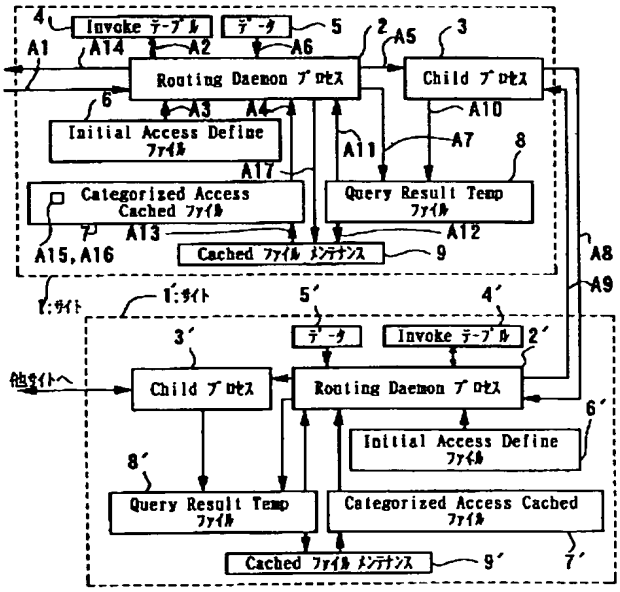
【図３】　本発明における該当データのマイニング方式の手順を示すフローチャート。

【図４】　従来方式の概要図。

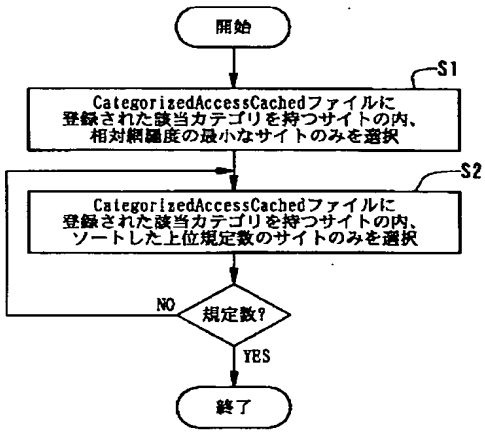
【符号の説明】

１　データ転送装置（サイト）
２　Routing Daemon プロセス
３　Child プロセス
４　Invoke テーブル
５　データ
６　Initial Access Define ファイル
７　Categorized Access Cached ファイル
８　Query Result Temp ファイル
９　Cached ファイル メンテナンス
１１　Httpd プロセス
１２　コンテントルータ
１３　ルーティングデータベース
１４　Refinement データベース
１５　コンテンツラベル
１６　サーチモジュール
１７　WAIS プロセス
１８　WAIS データベース

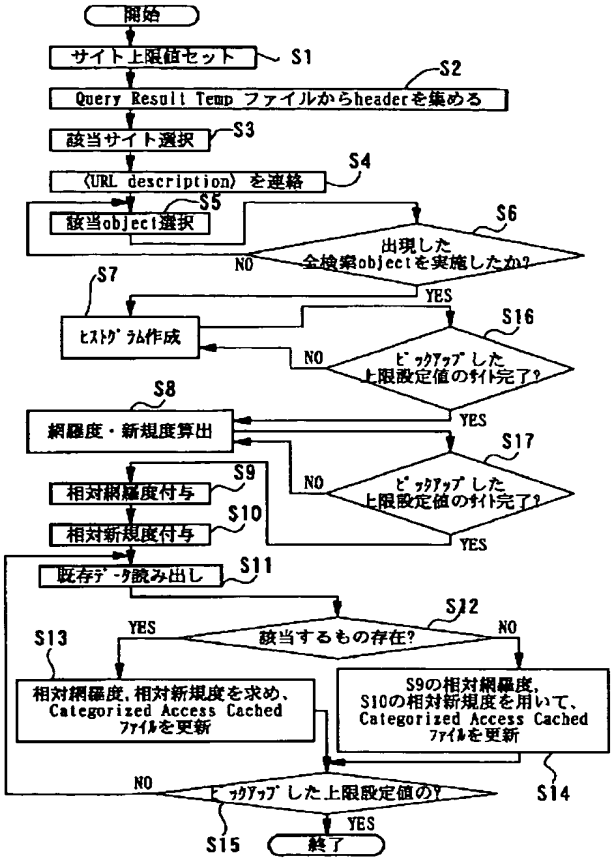
【図１】



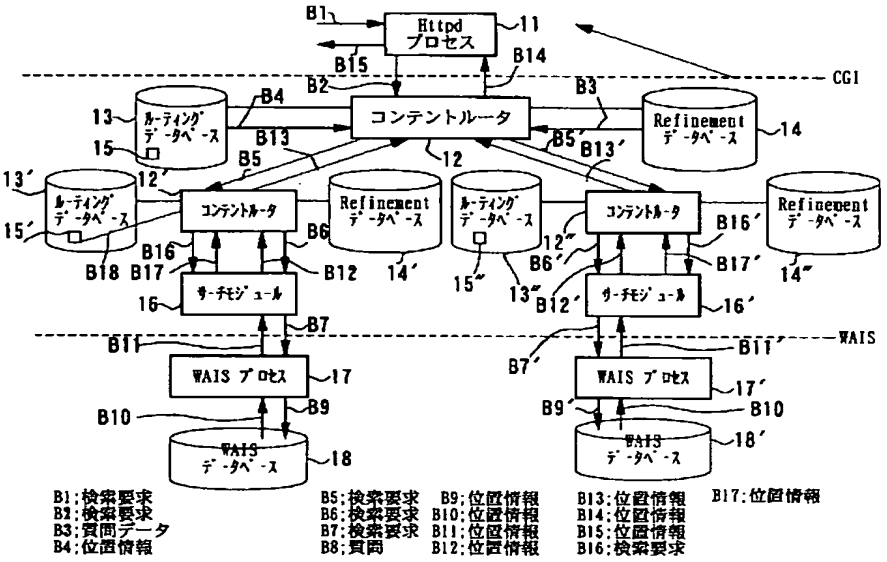
【図２】



【図３】


開始
サイト上限値セット
S1
Query Result Temp ファイルからheaderを集める
S2
該当サイト選択
S3
〈URL description〉を連絡
S4
該当object選択
S5
出現した全検索objectを実施したか?
S6
NO
YES
ヒストグラム作成
S7
ピックアップした上限設定値のサイト完了?
S16
NO
YES
網羅度・新規度算出
S8
相対網羅度付与
S9
相対新規度付与
S10
ピックアップした上限設定値のサイト完了?
S17
NO
YES
既存データ読み出し
S11
該当するもの存在?
S12
YES
NO
S13
相対網羅度,相対新規度を求め、Categorized Access Cached ファイルを更新
S9の相対網羅度,S10の相対新規度を用いて、Categorized Access Cached ファイルを更新
S14
ピックアップした上限設定値の?
NO
YES
S15
終了

【図４】


B1
Httpd プロセス
11
B15
B14
B2
CGI
B3
B4
13
ルーティング データベース
15
コンテントルータ
Refinement データベース
14
B13
B5
B5′
B13′
12
13′
12′
15′
B16
B6
B18
B17
B12
14′
16
サーチモジュール
B7
B11
WAIS
WAIS プロセス
17
B10
B9
WAIS データベース
18
12″
15″
13″
B6′
B12′
B16′
B17′
14″
16′
B7′
B11′
17′
B9′
B10′
18′
B1;検索要求
B2;検索要求
B3;質問データ
B4;位置情報
B5;検索要求
B6;検索要求
B7;検索要求
B8;質問
B9;位置情報
B10;位置情報
B11;位置情報
B12;位置情報
B13;位置情報
B14;位置情報
B15;位置情報
B16;検索要求
B17;位置情報